Quefrency Approach to Audio Deepfake Detection

Kanishq Singhal, Aditya Goyal and Priyanka Gupta
The LNM Institute of Information Technology, Jaipur, India
E-mail: 23ucs605@lnmiit.ac.in, 23ucs512@lnmiit.ac.in, priyanka.gupta@lnmiit.ac.in

***Abstract*—This work aims at audio deepfake detection (ADD) in three different scenarios-only deepfake, re-recorded-deepfake, and replayed-deepfake. To that effect, the existing Fake or Real (FoR) dataset is enhanced by a reverberation topology leading to a novel version of the dataset (called as rev-FoR-rerec) to simulate replayed-deepfakes in varying acoustic environments. Furthermore, this work investigates the significance of quefrency-based representation for ADD. Our findings indicate that as compared to mel-spectrogram representations, the proposed quefrency-based system achieves improved EERs on *both* testing and validation sets of all three attack scenarios, except for the case of replayed-deepfake where the proposed system performs nearly equal to the mel-spectrogram with an absolute difference in testing EER as 0 and an absolute difference in validation EER as 0.09. Furthermore, our sub-band analysis shows that the bands 0-1000 Hz and 7000-8000 Hz are the most discriminating subbands for ADD. Experiments are also performed to investigate the effect of room size and reverberation damping factor on the detection of replayed-deepfakes.**

I. Introduction

Deepfakes pose a significant threat due to their increasing realism and potential for manipulation in various media formats like text, audio, images, and videos [1], [2]. This has detrimental implications for privacy, security, and societal integrity [3]. Audio deepfakes, created through AI-generated or manipulated audio, aim to mimic a target speaker's voice convincingly. Advancements in text-to-speech (TTS) and voice conversion (VC) techniques have facilitated the production of high-quality audio deepfakes [4], [5].

Detecting audio deepfake is a challenging task that involves various techniques from signal processing, machine learning, and deep learning [6], [7]. Previous research has shown promising results [8]–[10]. Many existing works focus on summarizing past spoofing attacks, or protecting automatic speaker verification (ASV) systems [11], [12]. However, datasets released in the ASVSpoof challenges lack usability for research because of undisclosed synthesis algorithms due to the nature of the ASVSpoof challenges. To that effect, we have utilized the Fake-Or-Real dataset for ADD in this work.

So far, most of the neural network architectures for ADD utilize spectrogram analysis [13]–[15], however, quefrency-based representations such as cepstrograms are less explored. To that effect, this work investigates the significance of quefrency approach to ADD. In addition, since TTS and VC technologies have eased the generation of audio deepfakes, deepfake attacks, like the replay attack have become easier to mount [16]. Hence, detecting these two attacks is crucial to ensure security and integrity. To that effect, a scenario of replayed-deepfake detection is also introduced in this work. In particular this paper has the following contributions:

- Quefrency-based features are proposed for ADD, and hence cepstral vs. spectral analysis is done to investigate which domain is more suited for ADD.
- Apart from the scenario where a fake/spoofed utterance is synthesized using deepfake only, we have considered two more scenarios- *rerec-deepfake* and *replayed-deepfake*. Rerec-deepfake considers the situation when the deepfaked utterance is transmitted through a communication channel (like a phone call or voice message). Replayed-deepfake considers the scenario where the deefaked signal is recorded and played back like a replay attack. Given that replay attacks are the easiest to execute from an attacker's perspective [17], it becomes important to evaluate countermeasure systems for replayed-deepfake scenario as well, thereby attempting towards generalization of countermeasure systems.
- To simulate the *replayed-deepfake* scenario, a new dataset (namely rev-FoR-rerec) is created where the deepfaked audio is recorded and replayed in rooms of varying size, with reverberations of varying intensities.
- To that effect, a reverberation topology for simulating the replayed-deepfake scenario is also proposed in this work.
- A lesser-known dataset *Fake-or-Real (FoR)* dataset is used in this work. Unlike the ASVSpoof 2021 DF dataset, this dataset is balanced in terms of gender and class, and normalized in terms of volume and number of channels. It should also be noted that the testing set of the FoR dataset consists of utterances generated using Deepvoice3, Baidu TTS, Amazon AWS Polly and Cloud, WaveNet, which are Google text-to-speech algorithm renowned for producing speech that closely resembles human voices hence testing our model on real-world conditions.
- Analysis on the effect of reverberation parameters affecting room-size and reverberation intensity has also been done the on *replayed-deepfake* scenario.
- Experiments are also performed on the ASVSpoof 2019 Logical Access (LA) dataset, to investigate the performance of the proposed ADD system on LA scenarios.

II. Proposed Work

This work presents three different attacking scenarios. In Fig. 1 (A), deepfake speech is generated using spoofing algorithms such as TTS (Text-to-Speech) or VC (Voice Conversion). In realistic scenarios, as shown in Fig. 1 (B), an attacker can re-record a deepfake utterance using a device like a smartphone and send it to a victim's smartphone through

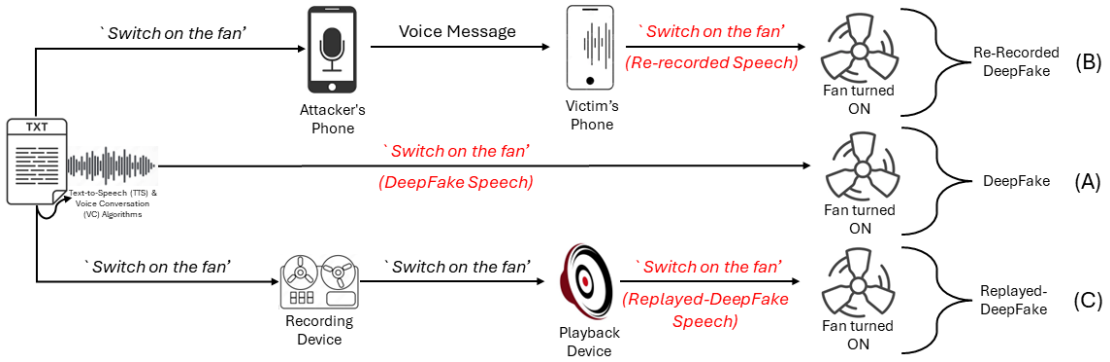

Fig. 1. Different types of attack scenarios: (A) Only Deepfake, (B) Re-Recorded Deepfake, and (C) Replayed-Deepfake

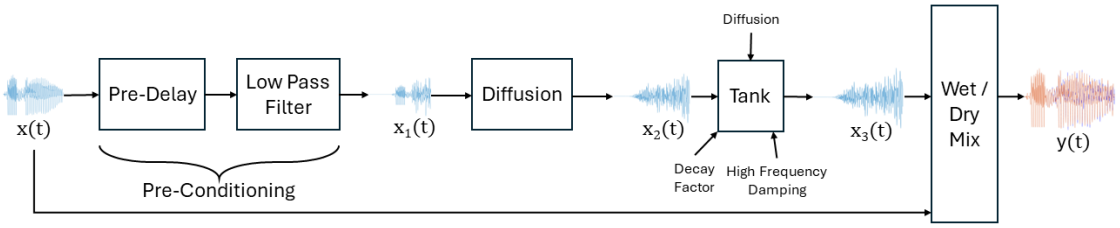

Fig. 2. Reverberated Topology for Replayed-Deepfake Scenario

a voice message or communication channel. Additionally, as illustrated in Fig. 1 (C), we introduced another potential attacking strategy that involves recording and playing back deepfake audio, which we refer to as *"replayed-deepfake"* in this work.

A. Proposed replayed-deepfake with reverberation topology

To further expand the scope of the dataset for anti-spoofing research, a novel "Reverberated FoR Re-Recorded (rev-for-rerec)" dataset is designed in this work. To bridge the gap in the existing ADD dataset we incorporated synthetic reverberations of varying specifications (such as room size, high frequency damping and pre-delay), thereby offering a more comprehensive evaluation platform for anti-spoofing models. This additional layer of intricacy to the dataset, which simulates real-world scenarios, reflects the diverse environments and conditions in which spoofing attacks might occur better. To that effect, we have introduced reverberation in the existing Fake or Real dataset using the reverberation.

The reverberation topology [18] as shown in Fig. 2 begins with *pre-conditioning*. Preconditioning comprises of two stages: 1) Pre-delay, and 2) Low-pass filtering. *Pre-delay* is the time between hearing direct sound and the first early reflection and is expressed as:

$$x_p(t) = x(t-k) \tag{1}$$

where k is the pre-delay constant. A smaller value of k is preferred for short-duration utterances because it ensures that the reverberation effect occurs sooner. Now, the pre-delayed signal is passed through the *low-pass filter* expressed as:

$$LP(z) = \frac{1-\alpha}{1-\alpha z^{-1}} \tag{2}$$

where α is estimated as:

$$\alpha_1 = \exp\left(-2\pi \times \frac{f_c}{f_s}\right) \tag{3}$$

where f_c is the *high cut frequency* and f_s is the sampling frequency. After the preconditioning, *diffusion* is performed which controls the density of the reverb tail and is specified as a real positive scalar in the range [0, 1]. To set the diffusion, the preconditioned signal, as shown in step 3 of Algorithm 1, is passed through a cascade of 4 *all-pass filters*, where each filter is expressed as:

$$AP(z) = \frac{\beta + z^{-k}}{1 + \beta z^{-1}} \tag{4}$$

where β is the diffusion constant and k is the delay. Increasing the diffusion pushes the reflections closer together, thickening the sound, whereas reducing it creates more discrete echoes. The filtered output is then fed into a *tank* to adjust the decay factor and high-frequency damping of the reverberation tail. The *high-frequency damping* adjusts the fading of high-frequency reflections, whereas the *decay factor* of the reverb tail determines how quickly the reflections lose energy, providing a longer reverb tail without overwhelming the original signal [18]. For this effect, the resultant signal is passed through a series of filters as demonstrated in step 4 of Algorithm 1. The modulated all-pass filter is:

$$ModulatedAP(z) = \frac{-\beta + z^{-k}}{1 - \beta z^{-1}} \tag{5}$$

where β is the diffusion constant from Eq. 4. Here, in the tank, the low-pass filter constant α (from Eq. 2) is used as the high-frequency damping constant. Finally, "*wet-dry mix*" is performed on the signal. The wet-dry mix ratio W (where, $W \in [0, 1]$) controls the ratio of the reverberated signal to the original, as illustrated in step 5 of Algorithm 1.

B. Proposed Quefrency-based representation for ADD

Most of the work on ADD has been done by analyzing the signal using spectrograms in linear or mel-scale [13]–[15]. However, quefrency-based representations such as cepstro-

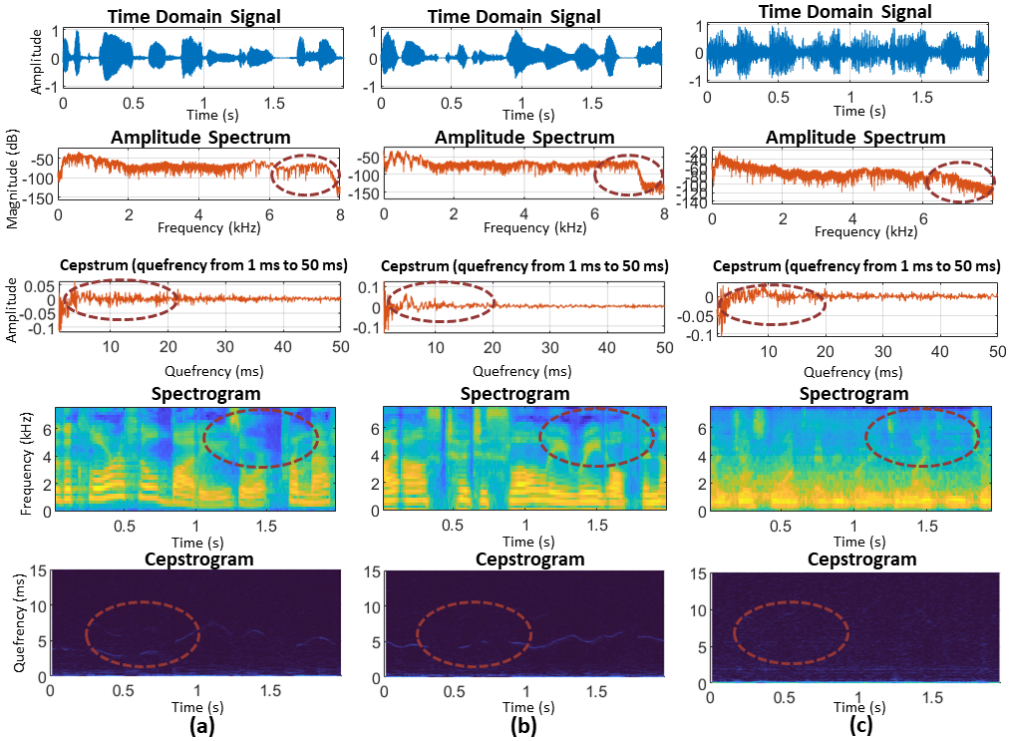

Fig. 3. Analysis of speech signals using various representations. (a) Genuine utterance, (b) Only Deepfake, (c) Replayed-Deepfake

Algorithm 1: Proposed Reverberation Algorithm for replayed-deepfake attack

Input: Input: Speech signal $x(t)$

Output: Output: Reverberated Speech $y(t)$

$x_p(t) \leftarrow Pre-Delay\ [x(t)]$

$x_1(t) \leftarrow LowPass\ Filter\ [x_p(t)]$ } *Pre-Conditioning*

$x_2(t) \leftarrow AP_i\ [x_1(t)]\ where\ AP_i$ is the i^{th} All-Pass Filter and $i = 1, 2, 3, 4$

$x_3(t) \leftarrow AllPassFilter \leftarrow LowPassFilter \leftarrow Modulated\ AllPassFilter\ [x_2(t)]$

$y(t) = (1-k)x(t) + Wx_3(t)$

grams are less explored. These representations have been used in pitch detection [19], echo removal [20], and noise reduction [21], [22]. Since deepfake algorithms are known to introduce artifacts, and the replayed deepfakes have reverberation, this work investigates the importance of quefrency-based representations on ADD.

Quefrency-domain representations measure cepstrum of a signal, which comes from the representation of a homomorphic system for convolution [23]. The word *cepstrum* is called "spectrum of the spectrum", and is estimated as [23] :

$$C = \left| \mathcal{F}^{-1} \left\{ \log \left(|\mathcal{F}\{x(t)\}|^2 \right) \right\} \right|^2 \tag{6}$$

where $x(t)$ is a signal, and $\mathcal{F}$ is its Fourier transformation.

A visual representation of the cepstrum over time is known as *cepstrogram*. It allows one to observe changes in the cepstral features (*rahmonics*) of a speech over time, similar to how a spectrogram shows changes in the spectral content of a speech.

C. *Frequency* vs *Quefrency Analysis*

To analyze the cases of genuine, only deepfake and replayed-deepfake utterances, Fig. 3 shows the differences between frequency and quefrency representations. In particular, Fig. 3 (a) shows the genuine utterance, Fig. 3 (b) shows the only deepfake utterance followed by Fig. 3 (c) which shows replayed-deepfake utterance. On analyzing the amplitude spectrums of the utterances, we observe that in Fig. 3 (a), the trend of the amplitude spectrum is *nearly* constant throughout the frequency range of 6-8 kHz as shown using a dotted circle. However, in Fig. 3 (b), we see that the amplitude spectrum is nearly constant *only* till 7 kHz, and then a sudden drop is observed, as marked by the circle. Meanwhile in Fig. 3 (c), the amplitude spectrum decreases gradually from 6 kHz onwards which is marked using a circle. Furthermore, we also analyze cepstrums of the utterances with quefrency from the range of 1 ms to 50 ms. The quefrency in the range of 1 ms to 50 ms is chosen because most of the variations have been observed in this range. The cepstrum has the highest distortion in Fig. 3 (c) followed by Fig. 3 (a) and Fig. 3 (b) respectively as marked by the circle. This indicates that the cepstrum for *only deepfake* (Fig. 3 (b)) fails to capture the environmental artifacts unlike the cepstrum in Fig. 3 (a) and Fig. 3 (c). This further indicates that the deepfake algorithms are not able to generate *naturalness* like genuine speech.

On analyzing the utterances using spectrogram, we observed that in genuine utterance (Fig. 3 (a)), the energy fading appears natural without any abrupt boundaries as opposed to the case of only deepfake utterance in Fig. 3 (b), where no continuous band of energy is observed as indicated by the circle. Furthermore, we observe that in replayed-deepfake

(Fig. 3 (c)) utterance, there is continuous energy in the region marked by the circle in the figure. This indicates that the genuine speech does not have as much energy as compared to the speech generated due to reverberation. However, when we analyze the utterances using quefrency-based representation or cepstrogram we observe that it is easier to differentiate between the genuine and replayed-deepfake speech than the genuine and only deepfake utterance because more dominant *rahmonic* peaks are observed in genuine speech as compared to replayed-deepfake speech which is marked with the help of the dotted circle.

III. Experimental Setup

A. Datasets

1) Fake or Real (FoR): The Fake or Real (FoR) has 4 versions namely, original, normalized, 2-sec version, and a re-recordered version [24]. In this work, experiments have been performed on the FoR 2-sec and For re-rec datasets. Both of these versions are balanced in terms of the number of utterances in each class, and gender variability of speakers. The FoR 2-sec and FoR re-rec versions correspond to deepfake only, and recorded-deefake attack scenarios, respectively, as discussed in Section II. The partitioning details of both of these datasets are identical [24].

2) Proposed Rev-rerec dataset: The replayed-deepfake attack scenario is emulated by incorporating reverberation into the for-rerec dataset, as discussed in Section II. In particular, for our experiments, we have fixed pre-delay factor at 0.25, diffusion as 0.5, and wet/dry mix at 0.3. The reverberation parameters, like decay factor and high-frequency damping, have been varied to investigate their effect on the performance.

B. Classifier and Performance Metrics

The CNN model used in this work consists of three convolutional blocks: Convolution 1, Convolution 2, and Convolution 3 where each convolutional block includes a 2-D convolutional layer with a kernel size of 3x3, stride of 1 and padding of 1, Rectified Linear Activation (ReLU) layer, and a max-pooling layer with a kernel size of 2x2. The output feature maps from the convolutional blocks are processed by two fully connected layers with 128 hidden units. This layer flattens the feature maps and then passes them through a ReLU activation function. The model is trained using the Adam optimizer with a learning rate of 0.001, and the loss is calculated using cross-entropy loss. The training process runs for 20 epochs, with a batch size of 32. The performance of the model is evaluated using Accuracy and Equal Error Rate (EER).

C. Baseline

In this work, we have considered the baseline from [10]. We investigate the significance of quefrency-based representation and mel-spectrogram. Both the representations are extracted using 30 ms of window length, with a window shift of 15 ms.

IV. Experimental Results & Analysis

A. Comparing Proposed System with the Existing Works

Experiments were performed using the proposed quefrency-based features and compared with the baseline [10] as shown in Table I. It can be observed that the proposed quefrency-based feature outperforms the existing baseline system on all the attack scenarios and datasets. In particular, we observed an absolute increase of 16.64%, 60.05%, 39.70%, and 8.63% in the testing and validation sets of the for-2sec and for-rerec datasets, respectively.

TABLE I
Existing System vs Proposed System

Dataset Used	Existing System [10]		Proposed System	
	Testing	Validation	Testing	Validation
FoR-2sec	71.14	38.29	**87.78**	**98.34**
FoR-rerec	45.96	86.47	**85.66**	**95.10**
rev-FoR-rerec	-	-	**100**	**99.87**

B. Feature Wise Comparison

To compare the performance of quefrency-based representation with a mel-spectrogram, experiments were performed on all the three attack scenarios, as shown in Table II. It can be seen that the performance of quefrency-based representation is better than the mel-spectrogram in most cases except in the rev-FoR-rerec dataset. However, it should be noted that only the best performances are shown in this table, and the detailed comparison is shown in Section IV-D. To further compare the features, experiments were performed on the widely known ASVspoof 2019 LA dataset. The quefrency-based feature achieved an accuracy of 89.63%, while the mel-spectrogram attained an accuracy of 89.60%, indicating no significant difference in performance.

TABLE II
Feature Wise Comparison: (A) Cepstrogram, (B) Mel Spectrogram

Attack Scenario		Only Deepfake		Rerecorded-Deepfake		Replayed-Deepfake	
Dataset Used		FoR-2sec		FoR-rerec		rev-FoR-rerec	
Feature Used		A	B	A	B	A	B
Accuracy (%)	Testing	**87.78**	87.13	**85.66**	81.5	**100**	100
	Validation	98.34	**98.44**	95.10	**96.17**	99.82	**99.87**
EER (%)	Testing	12.5	**8.64**	**13.24**	16.19	**0**	0
	Validation	1.70	**1.56**	**4.11**	4.64	0.26	**0.17**

C. Analysis on Only-Deepfake Scenario: Effect of Sub-Band Frequency Ranges

To observe the effect of the frequency range for deepfake detection, the full spectrum of the utterances in the FoR 2-sec dataset was divided into 8 sub-bands of bandwidth of 1000 Hz each (given the maximum frequency is 8000 Hz).

TABLE III
Effect of Sub-Band Frequency Ranges on ADD

Subband (in Hz)	Accuracy (%)		EER (%)	
	Testing	Validation	Testing	Validation
0 - 1000	86.40	**98.20**	13.05	**1.84**
1000 - 2000	64.98	83.69	33.09	14.79
2000 - 3000	64.15	87.44	32.90	14.79
3000 - 4000	50.92	90.34	46.88	9.77
4000 - 5000	61.76	88.50	34.74	10.62
5000 - 6000	49.45	88.50	51.47	11.46
6000 - 7000	66.27	89.67	34.01	9.98
7000 - 8000	**95.31**	97.38	**2.21**	2.76

TABLE IV
EFFECT OF REVERBERATION PARAMETERS ON REPLAYED-DEEPFAKE DETECTION

Decay Factor	High Frequency Damping	Mel-Spectrogram				Proposed Quefrency-based Representation			
		Accuracy (%)		EER (%)		Accuracy (%)		EER (%)	
		Testing	Validation	Testing	Validation	Testing	Validation	Testing	Validation
0.9	0	92.28	96.75	6.37	3.32	**97.67**	**98.08**	**1.72**	**1.92**
	0.25	90.56	96.43	9.56	3.50	**98.04**	**97.33**	**1.47**	**2.62**
	0.5	91.79	96.79	7.84	2.80	**97.79**	**98.08**	**1.47**	**2.01**
	0.75	98.28	**98.80**	1.72	**1.31**	**99.26**	97.64	**0.74**	2.62
	1	92.03	96.26	7.84	3.67	**97.43**	**97.46**	**2.21**	**2.54**
0.5	0	90.44	96.66	9.31	3.32	**99.63**	**98.26**	**0**	**1.57**
	0.25	97.67	98.31	2.70	1.92	**99.63**	**98.80**	**0**	**1.22**
	0.5	95.83	**97.95**	2.21	**1.92**	**98.90**	96.83	**0.25**	2.62
	0.75	96.81	97.77	3.43	2.01	**99.75**	**98.35**	**0.25**	**1.75**
	1	86.89	**96.52**	11.03	3.41	**98.77**	96.30	**1.47**	**3.32**
0.1	0	**99.88**	**99.91**	0	**0.26**	99.39	99.38	**0**	0.61
	0.25	99.75	**99.82**	0	**0**	**100**	99.64	**0**	0.26
	0.5	100	**99.87**	0	**0.17**	**100**	**99.82**	**0**	**0.26**
	0.75	99.63	**99.82**	0	**0.09**	**100**	99.69	**0**	0.26
	1	89.46	95.94	10.78	4.02	**93.14**	**96.88**	**0.74**	**2.50**

To that effect, experiments were performed by extracting sub-band-wise mel-spectrograms which were fed to the CNN for classification. The obtained performances are shown in Table III. It can be observed that the best performances for both testing and validation sets are observed in the first and last frequency bands (i.e. 0-1000 Hz and 7000-8000 Hz). This indicates that the sub-bands in the extremes of the spectrum play a significant role in ADD. It should also be noted that in the case of Voice Liveness Detection (VLD), the discriminating features w.r.t. pop noise are present in 0-40 Hz, i.e., on the lower extreme of the frequency range [25], [26]. Therefore, our subband analysis for ADD coincides with subband analysis for the VLD problem for low-frequency subbands.

Furthermore, we also observe in Table III that the validation accuracy is greater than the testing accuracy in all the subbands. This discrepancy arises because the testing dataset consists of unseen utterances. Moreover, the testing set was generated using WaveNet- a Google TTS algorithm, which is well-known to generate human-like speech.

D. Analysis on Replayed-Deepfake Scenario: Effect of Reverberation Parameters

To investigate the impact of room size and high-frequency damping on replayed-deepfake detection, experiments were performed on the proposed rev-FoR-rerec dataset. To that effect, Table IV shows the performances using mel-spectrogram and proposed quefrency-based feature (cepstrogram). For each value of decay factor $\left(DecayFactor \propto \frac{1}{RoomSize}\right)$, the high-frequency damping is varied to adjust the high-frequency reflections. The general trend observed in Table IV shows that on the testing set, the quefrency-based feature *consistently* performs better than the mel-spectrogram in *both* accuracy and EER across varying decay factors and damping values. In particular, as room size increases, both the features show improved performances, with the quefrency-based feature frequently achieving near-perfect to perfect accuracy and very low EER of 99% to 100% and 0 to 1% respectively. This indicates that large rooms contribute to better replayed-deepfake detection yielding robust results. The possible reason for this is that in large rooms, the delay between each reflection is more, as the second wave has to travel a larger distance before getting reflected, which makes it easier to distinguish between the replayed-deepfake utterance and the genuine utterance.

In the case of the proposed quefrency-based feature, for all the values of decay factor, and at moderate to high levels of the high-frequency damping factor (0.25 to 0.75), we can observe a high accuracy range $\sim 98 - 100\%$ and low EER values of $\sim 0 - 2\%$ are maintained, across both testing and validation sets. In particular, at a decay factor of 0.1, i.e., a large room, a *near-perfect* performance is observed. Furthermore, if compared to the mel-spectrogram, the proposed quefrency-based representation consistently achieves high performance across all damping levels exhibiting slightly better consistency.

Interestingly, in most of the cases of Table IV, the opposite of the trend mentioned in Section IV-C can be noticed. This is because the training and testing sets share the same intensities of reverberation, i.e., the replay component, because of which this replay component becomes 'seen' even in the testing set, and hence we observe a better performance in the testing set.

V. SUMMARY AND CONCLUSIONS

This study addresses the detection of audio deepfakes in three different scenarios, including replayed-deepfakes. To simulate real-world replayed deepfake attacks, we introduce a novel reverberated version of the FoR dataset (rev-FoR-rerec) with varying acoustic environments through a proposed reverberation topology. The effectiveness of quefrency-based features is investigated for all three scenarios. Furthermore, subband-wise analysis is done for ADD, and reverberation analysis is done for replayed-deepfake detection. This work

demonstrates the superior effectiveness of quefrency-based representations (cepstrograms) over mel-spectrograms, especially in replayed-deepfake attack. The obtained results highlight the importance of room size and high-frequency damping, with larger rooms enhancing the discriminability between genuine and replayed-deepfake utterances. Subband analysis reveals that the subbands in the extreme ranges, i.e., 0-1000 Hz and 7000-8000 Hz are the most informative for deepfake detection, coinciding with the findings of VLD in literature, where subbands in the extremely low-frequency ranges (0 to 40 Hz) are discriminating. In the future, the efficiency of VLD systems can be tested on audio deepfakes.

REFERENCES

[1] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[2] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017.

[3] B. Nguyen, "A couple in canada were reportedly scammed out of $21,000 after getting a call from an ai-generated voice pretending to be their son.," *The New York Times*, March 2023 {Last Accessed date : 24^{th} June, 2024}.

[4] P. Kawa, M. Plata, M. Czuba, P. Szymański, and P. Syga, "Improved deepfake detection using whisper features," *arXiv preprint arXiv:2306.01428*, 2023.

[5] J.-w. Jung, H.-S. Heo, H. Tak, *et al.*, "Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks," in *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2022, pp. 6367–6371.

[6] A. M. Almars, "Deepfakes detection techniques using deep learning: A survey," *Journal of Computer and Communications*, vol. 9, no. 05, pp. 20–35, 2021.

[7] X. Wu, R. He, Z. Sun, and T. Tan, "A light cnn for deep face representation with noisy labels," *IEEE transactions on information forensics and security*, vol. 13, no. 11, pp. 2884–2896, 2018.

[8] G. Gupta, K. Raja, M. Gupta, T. Jan, S. T. Whiteside, and M. Prasad, "A comprehensive review of deepfake detection using advanced machine learning and fusion methods," *Electronics*, vol. 13, no. 1, p. 95, 2023.

[9] J. Yi, C. Wang, J. Tao, X. Zhang, C. Y. Zhang, and Y. Zhao, "Audio deepfake detection: A survey," *arXiv preprint arXiv:2308.14970*, 2023.

[10] R. A. M. Reimao, "Synthetic speech detection using deep neural networks," 2019.

[11] A. Mittal and M. Dua, "Automatic speaker verification systems and spoof detection techniques: Review and analysis," *International Journal of Speech Technology*, vol. 25, no. 1, pp. 105–134, 2022.

[12] M. Li, Y. Ahmadiadli, and X.-P. Zhang, "Audio anti-spoofing detection: A survey," *arXiv preprint arXiv:2404.13914*, 2024.

[13] L. Pham, P. Lam, T. Nguyen, H. Nguyen, and A. Schindler, "Deepfake audio detection using spectrogram-based feature and ensemble of deep learning models," *arXiv preprint arXiv:2407.01777*, 2024.

[14] J. Khochare, C. Joshi, B. Yenarkar, S. Suratkar, and F. Kazi, "A deep learning framework for audio deepfake detection," *Arabian Journal for Science and Engineering*, pp. 1–12, 2021.

[15] R. Anagha, A. Arya, V. H. Narayan, S. Abhishek, and T. Anjali, "Audio deepfake detection using deep learning," in *2023 12th International Conference on System Modeling & Advancement in Research Trends (SMART)*, IEEE, 2023, pp. 176–181.

[16] P. Gupta and H. A. Patil, "Linear frequency residual cepstral features for replay spoof detection on asvspoof 2019," in *30th European Signal Processing Conference (EUSIPCO)*, 2022, pp. 349–353.

[17] P. Gupta, H. A. Patil, and R. C. Guido, "Vulnerability issues in automatic speaker verification (asv) systems," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2024, no. 1, p. 10, 2024.

[18] J. Dattorro, "Effect design, part 1: Reverberator and other filters," *Journal of the Audio Engineering Society*, vol. 45, no. 9, pp. 660–684, 1997.

[19] A. V. Oppenheim and R. W. Schafer, "From frequency to quefrency: A history of the cepstrum," *IEEE signal processing Magazine*, vol. 21, no. 5, pp. 95–106, 2004.

[20] L. R. Rabiner, R. W. Schafer, *et al.*, "Introduction to digital speech processing," *Foundations and Trends® in Signal Processing*, vol. 1, no. 1–2, pp. 1–194, 2007.

[21] O. B. Osman and M. H. Arbab, "Mitigating the effects of granular scattering using cepstrum analysis in terahertz time-domain spectral imaging," *PLoS One*, vol. 14, no. 5, e0216952, 2019.

[22] X. Miao, M. Sun, X. Zhang, and Y. Wang, "Noise-robust voice conversion using high-quefrency boosting via sub-band cepstrum conversion and fusion," *Applied Sciences*, vol. 10, no. 1, p. 151, 2019.

[23] B. P. Bogert, "The quefrency alanysis of time series for echoes: Cepstrum, pseudoautocovariance, cross-cepstrum and saphe cracking," in *Proc. Symposium Time Series Analysis, 1963*, 1963, pp. 209–243.

[24] R. Reimao and V. Tzerpos, "For: A dataset for synthetic speech detection," in *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, IEEE, 2019, pp. 1–10.

[25] P. Gupta and H. A. Patil, "Morse wavelet transform-based features for voice liveness detection," *Computer Speech & Language*, vol. 84, p. 101 571, 2024.

[26] P. Gupta, P. K. Chodingala, and H. A. Patil, "Morlet wavelet-based voice liveness detection using convolutional neural network," in *2022 30th European Signal Processing Conference (EUSIPCO)*, IEEE, 2022, pp. 100–104.